VOICE CONVERSION BASED ON SIMULTANEOUS MODELING OF SPECTRUM AND F0

Kaori Yutani, Yosuke Uto, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

Department of Computer Science and Engineering
Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

This paper proposes a simultaneous modeling of spectrum and F_0 for voice conversion based on MSD (Multi-Space Probability Distribution) models. As a conventional technique, a spectral conversion based on GMM (Gaussian Mixture Model) has been proposed. Although this technique converts spectral feature sequences nonlinearly based on GMM, F_0 sequences are usually converted by a simple linear function. This is because F_0 is undefined in unvoiced segments. To overcome this problem, we apply MSD models. The MSD-GMM allows to model continuous F_0 values in voiced frames and a discrete symbol representing unvoiced frames within an unified framework. Furthermore, the MSD-HMM is adopted to model long term correlations in F_0 sequences.

Index Terms— voice conversion, F_0 conversion, MSD-GMM, MSD-HMM

1. INTRODUCTION

Voice conversion is a technique for converting a certain speaker's voice into another speaker's voice. It can modify speech characteristics using conversion rules statistically extracted from a small amount of data . One of typical spectral conversion frameworks is based on a Gaussian Mixture Model (GMM) [1]. This method realizes a continuous mapping based on soft clustering. A more accurate formulation of spectral conversion based on ML (Maximum Likelihood) criterion has been presented [2]. In the ML-based conversion, both training and conversion process are consistently derived based on the single ML objective function.

In the conventional GMM-based method, spectral feature sequences are nonlinearly converted based on GMM. However, F_0 sequences are converted by a simple linear function. This is because F_0 is undefined in unvoiced segments; therefore F_0 sequences cannot be modeled by neither continuous nor discrete distributions. In the voice conversion system, there are four types of F_0 combinations of the source and target features ("voiced-voiced," "voiced-unvoiced," "unvoiced-voiced" and "unvoiced-unvoiced"). Although, a method which focuses only on "voiced-voiced" features has been proposed [3], this method may be insufficient as a statistical model for accurately representing whole F_0 sequences. In this paper, we propose a method for simultaneous modeling of spectrum and F_0 based on Multi-Space Probability Distribution (MSD) models [4]. In the proposed method, each feature of "voiced-voiced," "voiced-unvoiced," "unvoiced-voiced" and "unvoiced-unvoiced" is modeled in a different probabilistic space. Thus the proposed method can convert voiced segments into unvoiced segments and vice versa, if the spaces of "voiced-unvoiced" or "unvoiced-voiced" features are selected in the convertion process. In this paper, we use MSD-GMM and MSD-HMM as MSD models. The method based on MSD-GMM can convert F_0 nonlinearly. Furthermore, the MSD-HMM is adopted to model long time correlations in F_0 sequences.

The paper is organized as follows. Section 2 explains the conventional voice conversion technique based on GMM. The voice conversion techniques based on MSD-GMM and MSD-HMM are described in Section 3 and Section 4, respectively. Experimental results are reported in Section 5. Finally, conclusions are given in Section 6.

2. VOICE CONVERSION BASED ON GMM

To convert spectral features of a source speaker X to a target speaker Y, the joint probability density of two speaker's features are modeled by GMM. Let a vector $\boldsymbol{Z}_t = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ be a joint feature vector of the source one $\boldsymbol{X}_t$ and the target one $\boldsymbol{Y}_t$ at time t. In the GMM-based voice conversion, the vector sequence $\boldsymbol{Z} = \left[\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top, \ldots, \boldsymbol{Z}_T^\top\right]^\top$ is modeled by GMM $\boldsymbol{\lambda} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \mid i = 1, 2, \ldots, M\}$. The output probability of $\boldsymbol{Z}$ given GMM $\boldsymbol{\lambda}$ can be written as follows:

$$p\left(\boldsymbol{Z}|\boldsymbol{\lambda}\right) = \prod_t^T \sum_i^M w_i \mathcal{N}(\boldsymbol{Z}_t|\boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(Z)}) \tag{1}$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix}, \ \boldsymbol{\Sigma}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix} \tag{2}$$

where M is the number of mixtures, w_i is the mixture weight of the i-th component, $\boldsymbol{\mu}_i^{(\cdot)}$ and $\boldsymbol{\Sigma}_i^{(\cdot)}$ is the mean vector and covariance matrix, respectively.

2.1. Maximum likelihood spectral conversion

In the maximum likelihood spectral conversion [2], the optimal sequence of the target feature vectors $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top\right]^\top$ given a source feature vector sequence $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top\right]^\top$ is obtained by maximizing the following conditional distribution:

$$p\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}\right) = \prod_t^T \sum_i^M p(m_t = i|\boldsymbol{X}_t, \boldsymbol{\lambda}) p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_t = i, \boldsymbol{\lambda}) \tag{3}$$

where $\boldsymbol{m} = (m_1, m_2, \ldots, m_T)$ is a mixture index sequence. The conditional distribution given $\boldsymbol{X}$ also becomes a GMM, and its output probability distribution can be written as follows:

$$p\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_t = i, \boldsymbol{\lambda}\right) = \mathcal{N}(\boldsymbol{Y}_t|\boldsymbol{E}_i(t), \boldsymbol{D}_i) \tag{4}$$

and

$$\boldsymbol{E}_i(t) = \boldsymbol{\mu}_t^{(Y)} + \boldsymbol{\Sigma}_i^{(YX)} {\boldsymbol{\Sigma}_i^{(XX)}}^{-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_i^{(X)}\right) \tag{5}$$

$$\boldsymbol{D}_i = \boldsymbol{\Sigma}_i^{(YY)} - \boldsymbol{\Sigma}_i^{(YX)} {\boldsymbol{\Sigma}_i^{(XX)}}^{-1} \boldsymbol{\Sigma}_i^{(XY)} \tag{6}$$

Since equation (3) includes latent variables, the optimal sequence of $\boldsymbol{Y}$ is estimated via the EM algorithm. The EM algorithm is an iterative method for approximating the maximum likelihood estimation.

It maximizes the expectation of the complete data log-likelihood so called $\mathcal{Q}$-function (auxiliary function):

$$\mathcal{Q}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{\text{all } \boldsymbol{m}} p(\boldsymbol{Y}, \boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) \log p\left(\hat{\boldsymbol{Y}}, \boldsymbol{m}\middle|\boldsymbol{X}, \boldsymbol{\lambda}\right) \tag{7}$$

Taking the derivative of the $\mathcal{Q}$-function, the spectral sequence $\hat{\boldsymbol{Y}}$ which maximizes the $\mathcal{Q}$-function is given by

$$\hat{\boldsymbol{Y}} = \left(\overline{\boldsymbol{D}^{-1}}\right)^{-1} \overline{\boldsymbol{D}^{-1}\boldsymbol{E}} \tag{8}$$

where

$$\overline{\boldsymbol{D}^{-1}} = \text{diag}\left[\overline{\boldsymbol{D}_1^{-1}}, \overline{\boldsymbol{D}_2^{-1}}, \cdots, \overline{\boldsymbol{D}_T^{-1}}\right] \tag{9}$$

$$\overline{\boldsymbol{D}_t^{-1}} = \sum_{i=1}^{M} \gamma_i(t) \boldsymbol{D}_i^{-1} \tag{10}$$

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}} = \left[\overline{\boldsymbol{D}^{-1}\boldsymbol{E}_1}^{\top}, \overline{\boldsymbol{D}^{-1}\boldsymbol{E}_2}^{\top}, \cdots, \overline{\boldsymbol{D}^{-1}\boldsymbol{E}_T}^{\top}\right]^{\top} \tag{11}$$

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}_t} = \sum_{i=1}^{M} \gamma_i(t) \boldsymbol{D}_i^{-1} \boldsymbol{E}_i(t) \tag{12}$$

$$\gamma_i(t) = p(m_t = i|\boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}) \tag{13}$$

2.2. F_0 conversion

In the conventional method, F_0 is converted linearly using the following equation:

$$p_t^{(Y)} = \frac{p_t^{(X)} - \mu^{(X)}}{\sigma^{(X)}} \times \sigma^{(Y)} + \mu^{(Y)} \tag{14}$$

where $p_t^{(X)}$ and $p_t^{(Y)}$ are input and converted F_0 values, respectively, $\mu^{(\cdot)}$ and $\sigma^{(\cdot)}$ are the mean and the standard deviation of F_0, respectively.

3. VOICE CONVERSION BASED ON MSD-GMM

3.1. Feature modeling by MSD

We consider G spaces $(R^1, R^2, \cdots, R^G)$ shown in Fig. 1, which specified by space index $g = 1, 2, \cdots, G$, where R^g is n_g-dimensional space. Each space R^g has a probability density function $(\mathcal{N}_1^{n_1}, \mathcal{N}_2^{n_2}, \cdots, \mathcal{N}_G^{n_G})$ and its probability $(c_1, c_2, \cdots, c_G)$, where $\sum_{g=1}^{G} c_g = 1$. Each event E is represented by a random variable $\boldsymbol{o}$ which consists of a continuous random variable $\boldsymbol{x} \in R^n$ and a set of space indices $\boldsymbol{X}$, that is,

$$\boldsymbol{o} = (\boldsymbol{X}, \boldsymbol{x}) \tag{15}$$

where all spaces specified by $\boldsymbol{X}$ are n-dimensional. The observation probability of $\boldsymbol{o}$ is defined by

$$p(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} c_g \mathcal{N}_g^{n_g}(V(\boldsymbol{o})) \tag{16}$$

where $V(\boldsymbol{o}) = \boldsymbol{x}, S(\boldsymbol{o}) = \boldsymbol{X}$. We assume that R^g contains only one sample point if $n_g = 0$. Accordingly, letting $P(E)$ be the probability distribution, we have

$$\int p(\boldsymbol{o}) d\boldsymbol{o} = \sum_{g=1}^{G} c_g \int \mathcal{N}_g^{n_g} d\boldsymbol{x} = 1 \tag{17}$$

It noted that, although $\mathcal{N}_g^0(V(\boldsymbol{o}))$ does not exist since R^g contains only one sample point, for simplicity of notation, we defines as $\mathcal{N}_g^0(V(\boldsymbol{o})) = 1$.

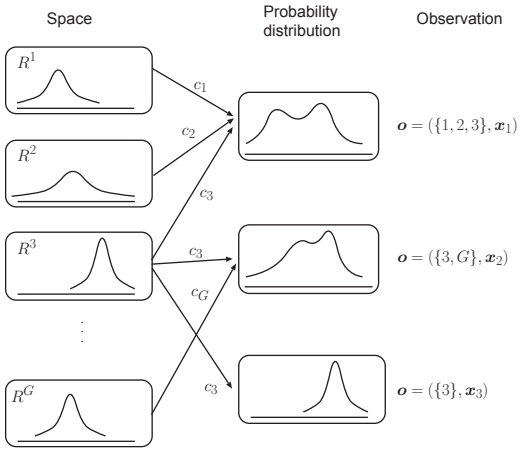

Fig. 1. MSD and observation vectors

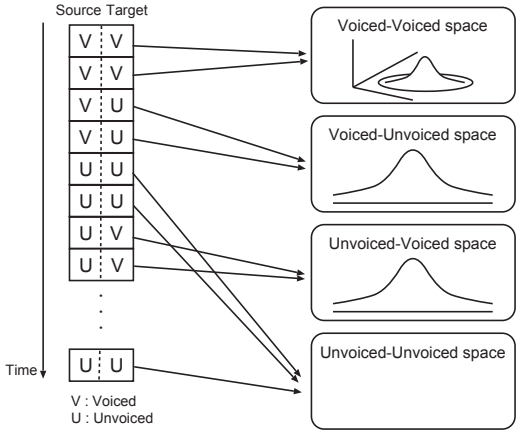

Fig. 2. F_0 modeling based on MSD

3.2. Modeling of spectrum and F_0

In the proposed method based on MSD models, a joint feature sequence $\boldsymbol{Z} = \left[\boldsymbol{Z}_1^{\top}, \boldsymbol{Z}_2^{\top}, \cdots, \boldsymbol{Z}_T^{\top}\right]^{\top}, \boldsymbol{Z}_t = \left[\boldsymbol{X}_t^{\top}, \boldsymbol{Y}_t^{\top}\right]^{\top}$ consists of spectral and F_0 feature vectors, where $\boldsymbol{X}_t = \left[\boldsymbol{c}_t^{(X)\top}, \Delta\boldsymbol{c}_t^{(X)\top}, p_t^{(X)\top}, \Delta p_t^{(X)\top}\right]^{\top}$ and $\boldsymbol{Y}_t = \left[\boldsymbol{c}_t^{(Y)\top}, \Delta\boldsymbol{c}_t^{(Y)\top}, p_t^{(Y)\top}, \Delta p_t^{(Y)\top}\right]^{\top}$ are a source and target vector, respectively. Each feature vector consists of spectrum feature $\boldsymbol{c}_i^{(\cdot)}$, F_0 feature $p_i^{(\cdot)}$ and their dynamic features denoted by $\Delta(\cdot)$. Fig. 2 shows F_0 modeling based on MSD. In the proposed method, there are four types of F_0 combination of source and target features ("voiced-voiced," "voiced-unvoiced," "unvoiced-voiced" and "unvoiced-unvoiced"). Each feature is modeled in a different probabilistic space by a single Gaussian distributions.

3.3. F_0 conversion

In the conversion process based on MSD, first, the converted sequences are determined whether these are voiced or unvoiced segments; if the input is "unvoiced" symbol, the posterior distribution are determined using the space weight. If the input is voiced feature, next, the values of each voiced segment are estimated. These F_0 values are generated from equation (8) similarly to the conventional spectral conversion.

4. VOICE CONVERSION BASED ON MSD-HMM

To perform modeling of long time correlations in F_0 sequences, MSD-HMMs are constructed which take account of phonetic contexts. However, context labels of input sequences are unknown in the conversion process, it should be estimated from input feature sequences. This means that conversion is performed based on one huge HMM in which context labels are regarded as latent variables. Fig. 3 shows the procedure of constructing one huge MSD-HMM. First, context dependent HMMs are constructed. Second, to overcome the overtraining problem, HMM states are shared by using a context clustering technique [5]. Furthermore, to model long time correlations more flexibly, sharing states along time are allowed as shown in the right figure of Fig. 4. Third, to regard contexts as hidden variables, one huge HMM is constructed by combining all HMMs dependently on contexts. Paths are added from the final states of HMMs to the initial states as shown in the right figure of Fig. 5. However, the computational complexity of model parameter re-estimation also becomes huge, because of the huge network of state transition. To overcome this problem, an HMM topology is minimized by assuming that states shared in the clustering are topologically identical as shown in the right figure of Fig. 6. However, it causes a problem that the state paths which do not exist before minimization are allowed.

5. EXPERIMENTS

5.1. Experimental conditions

Voice conversion experiments on the ATR Japanese database were conducted. We selected two sets of source and target speakers ("MTK→MHT," "MHO→MYI"). Each speaker uttered 503 sentences, and 450 sentences are used for training and remaining 53 sentences are used for evaluation. The speech data were downsampled from 20kHz to 16kHz, windowed at a 5-ms frame rate using a 25-ms Blackman window. Feature vectors consist of spectral and F_0 feature vectors. Each spectral feature vector consists of 24 mel-cepstral coefficients excepting the zeroth coefficient and their delta coefficients. Each F_0 feature vector consists of F_0 and its delta. The number of mixtures and states of HMM are varied among 32, 64, 128, 256 and 512.

In the experiment, the following four models were compared. "GMM": the conventional GMM-based method, "MSD-GMM": the proposed method based on MSD-GMM, "MSD-HMM1" and "MSD-HMM2": the proposed methods without and with minimization, respectively.

5.2. Objective evaluation

The mel-cepstral distortion (Mel-CD) was used as the objective measure of the spectral conversion. Voiced/unvoiced errors and F_0 distortion were also used as objective measures of F_0 conversion accuracy.

Fig. 9 and Fig. 10 show the results of subjective evaluation for "MTK→MHT" and "MHO→MYI," respectively. The proposed methods obtained a similar or slightly lower Mel-CD than the conventional method. Although, these differences are not large in terms of perception, it can be seen from the voiced/unvoiced errors that the proposed methods are smaller than the conventional method. This result shows the effectiveness of the voiced/unvoiced conversion based on MSD models. The F_0 distortion of "MTK→MHT" indicates that the proposed methods based on MSD models achieve higher performance than the conventional method. It is confirmed that the nonlinear conversion can convert F_0 accurately. In the result of "MHO→MYI," although the differences of F_0 distortion between

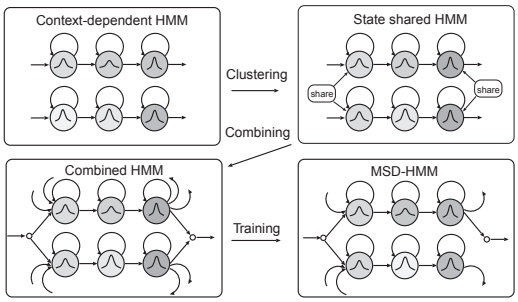

Fig. 3. The training process of MSD-HMM

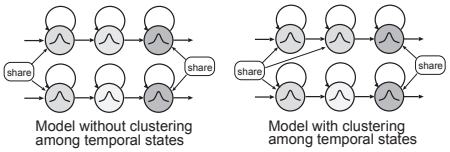

Fig. 4. The clustering among temporal states

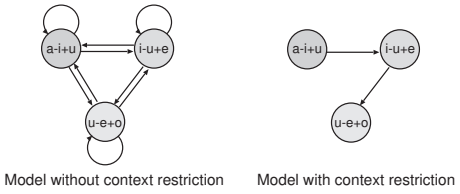

Fig. 5. The context restriction

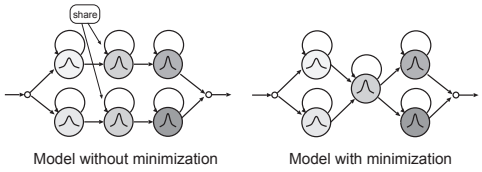

Fig. 6. The model structure minimization

the conventional and proposed methods is small when the number of states are small, improvements can be seen with increasing the number of states. However, comparing MSD-GMM and MSD-HMMs, no significant difference is observed in the objective evaluations.

5.3. Subjective evaluation

A DMOS test was performed for evaluating the similarity between the target and converted speech samples in speaker characteristics. The opinion score was set to a 5-point scale. Fifteen sentences were used for the evaluation set, and the number of listeners was 10. The number of mixtures/states are 256.

Fig. 7 and Fig. 8 show the results of the DMOS tests. Comparing the conventional method ("GMM") and the proposed methods ("MSD-GMM," "MSD-HMM1" and "MSD-HMM2"), the proposed methods are superior to the conventional method. This means that the nonlinear F_0 conversion and voiced/unvoiced conversion based on MSD models are effective for improving the similarity in the converted speech. However, comparing MSD-GMM and MSD-HMMs, MSD-HMMs are not superior to "MSD-GMM" as the objective evaluations. This might be because only the use of left and right phones

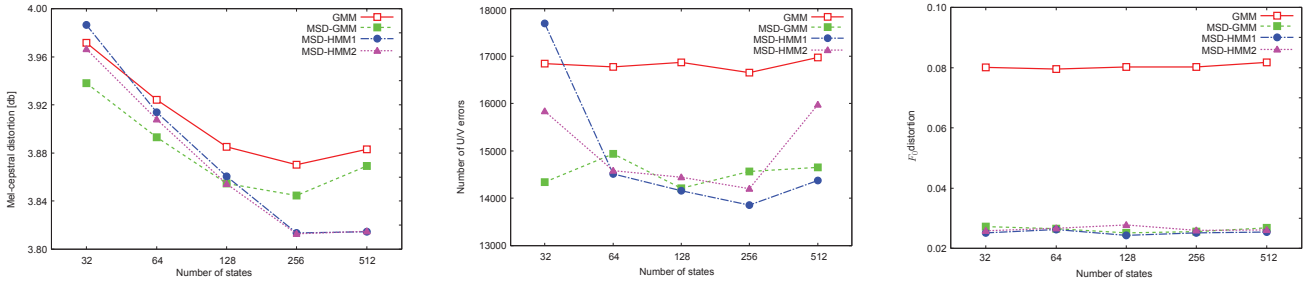

Fig. 9. Objective evaluation, "MTK→MHT" (left : Mel-CD, center : # of U/V errors, right : F_0 distortion)

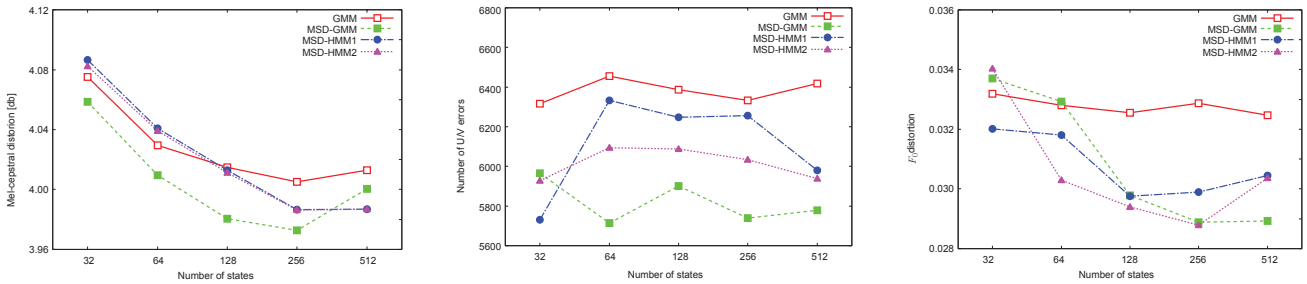

Fig. 10. Objective evaluation, "MHO→MYI" (left : Mel-CD, center : # of U/V errors, right : F_0 distortion)

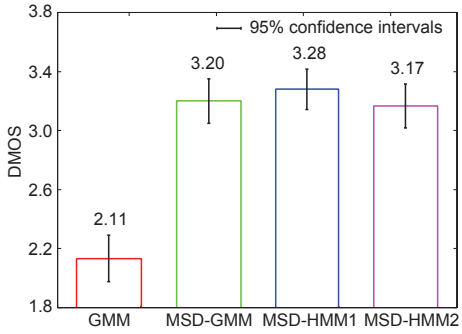

Fig. 7. DMOS (5-point scale), "MTK→MHT"

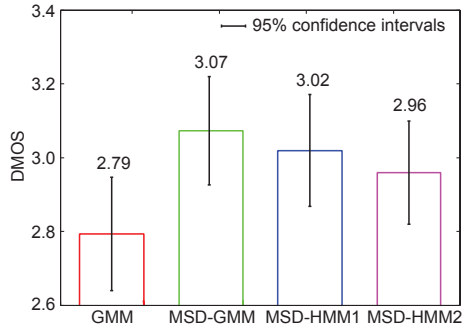

Fig. 8. DMOS (5-point scale), "MHO→MYI"

as contexts (triphone) is insufficient for modeling long time correlations of F_0 sequences.

6. CONCLUSION

This paper has proposed a simultaneous modeling technique of spectrum and F_0 for voice conversion. The proposed technique makes it possible to convert F_0 nonlinearly and to convert voiced segments into unvoiced ones and vice versa. In the experiments, it is confirmed that the proposed method achieved a higher performance than the conventional method.

7. REFERENCES

[1] Y. Stylianou, O. Cappe, E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *Proc. of IEEE Trans. Speech Audio Proc.*, vol.6, pp.131–142, Mar. 1998.

[2] T. Toda, A. W. Black, K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Trans. Audio Speech Language Proc.*, vol.15, pp.2222–2235, Nov. 2007.

[3] T. En-Najjary, O. Rosec, T. Chonavel, "A voice conversion method based on joint pitch and spectral envelope transformation," *Proc. of Interspeech*, pp.1225–1228, Oct. 2004.

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, " *Proc. of ICASSP*, vol.1, pp.229–232, May. 1999.

[5] J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD dissertation, Cambridge University, 1995.